Nederlandse samenvatting

Lexicaal-semantische relaties zijn relaties die betrekking hebben op woorden (het lexicon) al naar gelang hun betekenis (semantiek). Twee voorbeelden van lexicaal-semantische relaties zijn de synoniemrelatie en de co-hyponiemrelatie. Tussen de woorden *najaar* en *herfst* bestaat een synoniemrelatie, omdat de woorden dezelfde betekenis hebben. De woorden *bosbes* en *braam* staan ook in een lexicaal-semantische relatie. Beide woorden behoren tot dezelfde semantische klasse, namelijk de klasse van fruit. Woorden die tot dezelfde semantische klasse behoren worden co-hyponiemen genoemd.

Dit proefschrift behandelt twee onderzoeksvragen: ten eerste onderzoekt het verschillende methoden om lexicaal-semantische informatie automatisch uit grote tekstcorpora te extraheren. De centrale vraag daarbij is: welk type van lexicaal-semantische informatie resulteert uit de verschillende methoden? Ten tweede tracht het de verworven kennis toe te passen in een computerapplicatie. Hiermee proberen we te achterhalen, welk type lexicaal-semantische informatie waar inzetbaar is. De eerste vraagstelling is methodologisch van aard, terwijl de tweede toepassingsgericht is. Wij zullen het eerste doel nader toelichten en een korte samenvatting geven van onze bevindingen. Daarna zullen we het tweede doel behandelen.

Dit proefschrift behandelt drie methoden voor het automatische vergaren van lexicaal-semantische informatie. De drie methoden berusten op de *distributional hypothesis* 'distributionele hypothese' (Harris, 1968). De distributionele hypothese voorspelt dat overeenkomstige woorden in overeenkomstige contexten voorkomen (*similar words share similar contexts*). Dit impliceert dat we overeenkomstige woorden kunnen vinden door contexten van woorden te vergelijken. Wanneer wij, bijvoorbeeld, alle woorden selecteren die als lijdend voorwerp van het werkwoord *drinken* voorkomen, zien we meteen dat deze woorden tenminste één eigenschap overeenkomstig hebben, namelijk, het feit dat zij vloeibaar zijn. In dit voorbeeld is de context syntactisch van aard. We selecteren woorden in een bepaalde syntactische context, namelijk, het lijdend voorwerp van het werkwoord *drinken*.

Er zijn echter ook andere contexten denkbaar aan de hand waarvan de distributionele gelijkheid van woorden kan worden bepaald. We zouden, bijvoorbeeld, het woord *context* in brede zin op kunnen vatten en er meertalige parallelle corpora in kunnen betrekken, i.e. corpora die dezelfde inhoud beschrijven in verschillende talen. Door deze teksten automatisch te aligneren op woordbasis, kan de vertaling van woorden in verschillende talen worden benaderd. De vertaling die een woord in andere talen krijgt wordt zodoende de meertalige context van een woord.

De laatste context die we in dit proefschrift gehanteerd hebben voor de automatische vergaring van lexicaal-semantische informatie is de op nabijheid gebaseerde context. Hierbij maken alle betekenisvolle woorden in een bepaalde zin deel uit van de context voor een woord uit die bepaalde zin.

De drie verschillende contexten: de syntactische context, de meertalige context en de op nabijheid gebaseerde context liggen aan de basis van de drie verschillende methoden.

Om terug te komen op de eerste onderzoeksvraag: het type lexicaal-semantische informatie dat resulteert uit de drie methoden is zeer verschillend.

Wij hebben de gerangschikte lijsten van overeenkomstige woorden, die ons systeem voor ieder woord in onze testset berekent, geëvalueerd met behulp van het Nederlandstalige gedeelte van de lexicale database EuroWordNet (Vossen, 1998): Dutch WordNet (DWN). Hieruit blijkt dat de syntactische methode de minimale waarde, verkregen door de gemiddelde afstand tussen willekeurige woorden uit DWN te berekenen, overschrijdt (0.77 versus 0.26).

De methode die gebaseerd is op syntactische contexten levert synoniemen op en nog meer co-hyponiemen, zoals de woorden *bosbes* en *braam*. Verder vinden we hyperniemen, woorden die betekenis van een ander woord omvatten, zoals *fruit* een hyperniem van *bosbes* is. Ook de tegenovergestelde relatie vinden wij: *bosbes* als hyponiem van *fruit*.

Op de eerste positie wordt in zo'n 20% van de gevallen een synoniem aangetroffen. Een percentage van 100% is irrealistisch, aangezien niet elk woord een synoniem heeft. Volgens onze berekeningen heeft gemiddeld zo'n 60% van de zefstandige naamwoorden in DWN één of meer synoniemen. Vervolgens is het percentage co-hyponiemen dat het systeem geeft op de eerste positie twee keer zo hoog als het percentage synoniemen. Dit is niet wenselijk, maar ook niet erg verwonderlijk, aangezien woorden uit dezelfde semantische klasse, zoals *braam* en *bosbes*, vaak in dezelfde syntactische contexten voorkomen, bijvoorbeeld, als lijdend voorwerp van *eten* en gemodificeerd door het bijvoeglijk naamwoord *gezond*.

De op vertaalrelaties gebaseerde methode resulteert in minder co-hyponiemen. Een woord als *braam* wordt nu eenmaal zelden vertaald met een woord voor *bos-*

bes. Het percentage synoniemen dat gevonden wordt met deze methode ligt dan ook veel hoger: zo'n 30%. De opzet van evaluaties m.b.v. proefpersonen is lastig aangezien resultaten erg afhankelijk zijn van de opzet. Desalniettemin blijkt uit een evaluatie met proefpersonen dat zo'n 35% van de paren die door DWN als niet-synoniem worden bestempeld door een meerderheid wel degelijk als synoniem wordt geclassificeerd. De op vertaalrelaties gebaseerde methode gaat wel gebukt onder 'vuile' data, aangezien de vertalingen verkregen worden door automatische woordalignering, die niet altijd juist zijn.

De methode die uitgaat van de woorden, waarmee een bepaald woord voorkomt in een zin, levert een heel ander type lexicaal-semantische relaties op, namelijk, associaties, zoals het woord *wijn* voor het woord *feestje*. Deze lossere soort van lexicaal-semantische informatie is het gevolg van het feit, dat de context die gebruikt is, namelijk, alle betekenisvolle woorden in de zin, ook minder gestructureerd is dan bijvoorbeeld de syntactische context. We evalueren het resultaat op de woordassociatienormen van De Deyne and Storms (2008).

Nu we de drie methoden hebben toegelicht, komen we terug op de tweede onderzoeksvraag: Welk type lexicaal-semantische informatie is het best toepasbaar in de verschillende componenten van een vraag-antwoord-systeem?

Voor het beantwoorden van deze vraag hebben we verschillende soorten lexico-semantische informatie toegepast op verschillende componenten van het systeem. We hebben het systeem vervolgens getest op de vragensets van de Cross Language Evaluation Forum (CLEF[10]). Het forum voorziet in een testbatterij voor vraag-antwoord-systemen in meerdere Europese talen.

Zo gebruikten we co-hyponiemen verkregen met de syntactische methode voor het component vraagclassificatie, waar een vraag wordt ingedeeld in een klasse aan de hand van het verwachte antwoord. Een vraag als *Waar werd Audrey Hepburn geboren* wordt geclassificeerd als een locatie-vraag, aangezien de vraag een locatie als antwoord verwacht. Ook gebruikten we verschillende soorten van lexicaal-semantische informatie voor het uitbreiden van de zoektermen in het component *information retrieval*, waar tekstpassages worden geselecteerd waarin wij het antwoord op de vraag hopen te vinden. Hetzelfde type informatie gebruikten we in de module die de antwoorden uit de geselecteerde tekstpassages extraheert. Als laatste gebruikten we gecategoriseerde eigennamen voor het vinden van antecedenten voor nominale constituenten in de off-line-component van het systeem. In de off-line-component worden feiten uit teksten in tabellen verzameld zodat, wanneer hierover een vraag wordt gesteld, het antwoord snel en gemakkelijk kan worden opgezocht.

Uit de resultaten kunnen we opmaken dat vooral de categoriseerde eigennamen, een bij-product van de syntactische methode, het vraag-antwoord systeem

[10] http://clef-qa.itc.it

verbeteren. Zij verhogen de scores voor de *information retrieval* component en zorgen ervoor dat er voor bepaalde type vragen, namelijk de definitie- en WH-vragen, aanzienlijk vaker het goede antwoord gevonden wordt. Co-hyponiemen zijn toepasbaar voor het uitbreiden van bepaalde semantische klassen in DWN voor vraagclassificatie. Ook wordt er een kleine verbetering behaald, wanneer associaties worden gebruikt voor het uitbreiden van zoektermen in de component *information retrieval*.

Kortom, welke soort van lexico-semantische informatie adequaat is, verschilt per module en per soort van vragen. Het feit dat de CLEF vragenset veel vragen bevat waarin informatie wordt gevraagd over entiteiten levert een bijdrage aan het succes van de toepassing van gecategoriseerde eigennamen in deze taak. Ook de tegenvallende resultaten voor de toepassing van andere soorten lexicale informatie, vooral in de componenten *information retrieval* en antwoord matching en selectie, zouden te wijten kunnen zijn aan de specifieke opzet van de CLEF vragensets, waarin de lexicale variatie niet erg groot lijkt. De behaalde resultaten wijzen in een richting, maar kunnen niet worden gegeneraliseerd en zodoende als geldend gehouden worden voor vraag-antwoord-systemen in het algemeen, laat staan voor andere applicaties in de natuurlijke taalverwerking.